MongoDB-Hadoop Distributed and Scalable Framework for Spatio-Temporal Hazardous Materials Data Warehousing

Azedine Boulmakoul, Lamia Karim, Mohamed Haitam Laarabi, Roberto Sacile, Emmanuel Garbolino

MongoDB-Hadoop Distributed and Scalable Framework for Spatio-Temporal Hazardous Materials Data Warehousing

Azedine Boulmakoul[1], **Lamia Karim[1], Mohamed Haitam Laarabi[2], Roberto Sacile[2], and Emmanuel Garbolino[3]**

[1]FSTM, Department of computer sciences, LIM/IDS Lab. Faculty of Sciences and Technologies of Mohammedia. B.P. 146 Mohammedia – Morocco.Affiliations (azedine.boulmakoul@gmail.com, lkarim.lkarim@gmail.com)

[2] Department of Computer Science, Bioengineering, Robotics and Systems Engineering, University of Genoa, Italy. (haitam.laarabi@unige.it, roberto.sacile@unige.it)

[3] Crisis Research Centre (CRC), Mines ParisTech, Sophia Antipolis, France (emmanuel.garbolino@mines-paristech.fr)

Abstract: Today the demand for Carriage of Dangerous Goods is experiencing significant increase on Moroccan market. Each year, more than 15 million tons of dangerous goods are transported by road in Morocco. The transport of dangerous goods is regulated by a legal framework in line with international standards; including the European Agreement concerning the International Carriage of Dangerous Goods by Road (ADR) came into effect in Morocco in June 2003. With the aim to facilitate the deployment of some ADR guidelines, this project offers essential IT solutions for its application at the regional and national scale. In this context, the project involves development of software components for calculating safer time-dependent routes, spatial analysis of Voronoi network and establishment of a decisional database to capture HAZMAT (Hazardous Materials) shipments and occurring incidents or accidents. The framework that we propose assumes three major software components. The first is dedicated to the processing of time-dependent routes that considers risk and traffic conditions. The second is developing the transport network partitioning using Voronoi graph diagrams. This component is used for the purposes of management interventions. Finally, the last component provides a NoSQL database for the storage of HAZMAT events and shipping data. Other supports components are provided for collecting and visualizing of data and spatio-temporal events related to HAZMAT. All given components are integrated into an interoperable software infrastructure respecting intelligent transport systems architecture. This infrastructure is distributed and based on a service-oriented architecture. It is also scalable by integration of MongoDB with Hadoop for large-scale distributed data processing. In this work, we also give an assessment of the performance, scalability and fault-tolerance of using MongoDB with Hadoop, towards the goal of identifying the right architecture and software environment for HAZMAT spatio-temporal data analytics.

Keywords*: Dangerous Goods Transport; Real-time Data Collection; Time-dependent shortest path; Fuzzy weighted graph; HAZMAT Space Time Path Data Warehouse.*

1. INTRODUCTION

Thanks to the easy-to-use web technologies for route planning (such Google Map), decision makers can easily obtain computed paths, by only defining the start and end point, in addition to one or more criteria such: travel time, total path length, estimated travel cost or/and estimated risk. In Dangerous Goods Transport (DGT), decision makers are typically interested in the fastest and safest path for delivering the goods to multiple destinations (Boulmakoul et al., 2001). Usually only static information, gathered from past feedback studies, is taken into account when computing these algorithms, whereas the travel time besides the risk rate over a road segment depends on its traffic density, which in turn depends on time. This dependence requires a full knowledge on real-time traffic density, population density, nearby HAZMAT, information about transported HAZMAT, weather condition (Tomasoni et al., 2010).

Fortunately, latest information and communications technologies allow gathering the full knowledge using sensors positioned in different places (Ashwini et al., 2014), namely strategic spots in road network, vehicles, and weather stations. These data can be collected through GPRS or others technologies and stored in large databases (Boulmakoul, et al., 2001). Decision Makers can mine into these valuable information, apply some statistical analysis, and even mathematical models allowing the prediction of the traffic evolution to a certain degree of accuracy (Garbolino et al., 2007).

In such conditions, a road network depends therefore on the time instant at which the HAZMAT carrier traverses the road segment. It is called time-dependent network. Consequently, a practical issue arises: Sensors cannot cover the entire network, since in real world only specific roads network are constantly monitored. As well, sensors are not embedded in all HAZMAT carriers and remaining road networks are supplied with static information. A third source of information that can be considered is the traffic simulation method. It is based on Multi-Agent architecture where each agent of the system is autonomous and endowed with mental state, intelligence and ability to make decision taking into consideration the state of the environment and feedbacks from other agents (Laarabi et al., 2013). This system model offers a certain degree of dynamicity allowing thereby mimicry of real-world scenarios. The data generated by simulation allows studying not only different real world cases but also accidents and extreme cases. It is even possible injecting a large number of virtual carrier of dangerous goods on the monitoring system, as a strategy for pre-expansion of carriers' operator fleet.

The cost functions that bind a road segment to the corresponding travel time should be dynamic since the environment state is itself dynamic: real-time traffic information, real-time DGT carriers' feedback and traffic forecasting. Then for modelling travel time, risk and other fuzzy values, we use a Zadeh's Fuzzy Sets in Zadeh 1965 (Laarabi et al., 2014). Indeed, Fuzzy logic offers a more realistic interpretation of travel time and risk fluctuation. The Figure 1 summarizes the main aspects that shape the Intelligent Transport System proposed for managing the shipment of dangerous goods. Starting from real-time monitoring and multi-agent simulation, to time-dependent route planning on a partitioned transportation network with Voronoï Diagram (Laarabi et al., 2014) and (A. Mabrouk et al., 2009). Hereinafter, we will introduce each of these aspects addressed above and summarized on the Figure 1.

2. DATA COLLECTION

The GPRS-based real-time data transmission is commonly used system for monitoring behaviours and states of remote and scattered entities. The goal is data collection, risk assessment, service quality enhancement, and/or decision making whether at strategic, tactical or operational level. It has been applied, for instance, in the monitoring of water quality (Ji et al., 2011) and out-of-hospital cardiac patients (Xu et al., 2008). Furthermore, real-time data collectors can produce tremendous quantities of data: the Southern California Integrated GPS Network, or GPS Earth Observation Network System in Japan, collect data once per second with a latency of less than a second. Tera bytes of measurements per year can be generated by these kinds of systems as reported by the authors in (Aydin et al., 2007). Consequently, we were required to design a system that deals efficiently with real-time collection and data transmission surge.

The proposed solution concern two different scopes, whilst the first scope is linked to the protocol of communication between the server and vehicles, the other is rather focusing on server side data processing.

Data is transmitted directly according to a TCP protocol using GPRS connection. Once the server receives a quantity of messages, it proceeds to verification, storage and validity acknowledgment. In the case of reception of invalid messages, the server orders the clients, sources of invalidity, to cease transmission (It requires human intervention). On the server side, a handle is associated with each socket. These handles are run in an event processing loop, which are instantiated and block the calling function while there are on-going asynchronous operations.

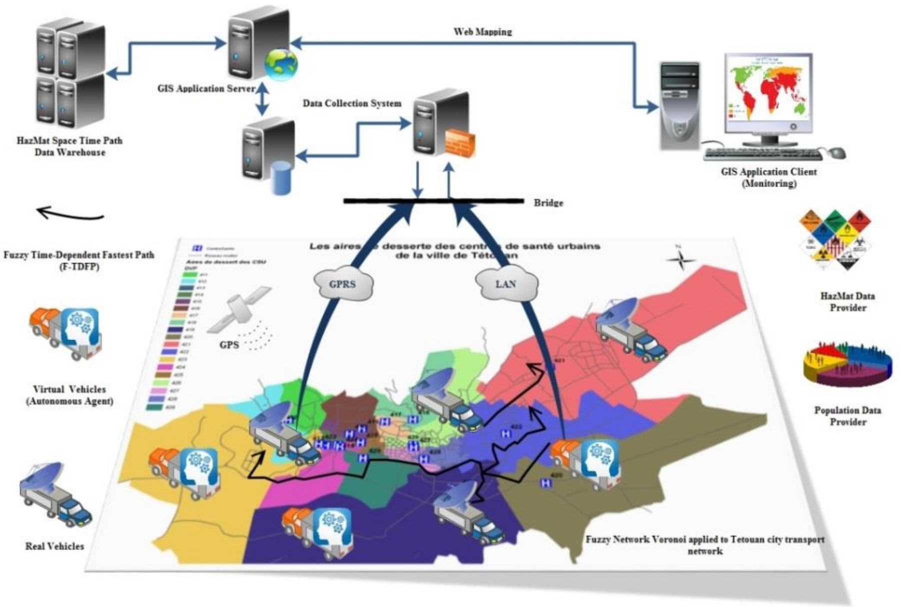

Figure 1. An Intelligent Transport System that includes DGT fleet monitoring, Data Collection, Virtual vehicles injection, GIS application, Database, Route planning considering travel-time and risk on a Fuzzy Voronoi Graph transport network

3. TIME-DEPENDENT FASTEST/SAFEST PATH SEARCH

Under reasonable assumptions, Dijkstra's algorithm solved the fastest path-finding problem on a time-dependent graph in polynomial time. However, in a large and detailed network, the computation of the Dijkstra's algorithm may require several seconds of CPU time, which is too long in a real-time system. Thereby, some speed-up techniques exist, such the use of heuristic strategies that impose search restriction on specifics road network and ignore all less significant ones. It is also possible to use mathematical models such the mixed-integer linear/nonlinear systems. In this work, we use a bidirectional A-star search based on restriction on forward search by bounding the set of nodes that will be explored by the backward search using a heuristic function. It has been proven as more efficient and less consuming in CPU time and memory than any other method does (Demiryurek et al., 2011).

The A-star algorithm uses two cost functions. Let considers v as an intermediate vertex within the $(s,d)-\mathrm{path}$. That is what makes $f(v) = g(v) + h(v)$:

- $g(v)$ The shortest travel-time of the $(s,v)-\mathrm{path}$
- $h(v)$ The estimated shortest travel-time of the $(v,d)-\mathrm{path}$

Therefore, by zeroing $h(v)$ or $g(v)$, A-star becomes Dijkstra's algorithm or Best First Search, respectively. The balancing of the two cost functions makes A-star a fast Point-to-Point shortest path algorithm. The goal is to propose the suitable $h(v)$ cost function, which will direct the search to the goal. Since $g(v)$ is the actual time-dependent fastest path (TDFP).

Let us consider s and d be respectively the source and destination nodes, and t_s the departure time at the source s. The time-dependent travel-time from s to d at t_s, which is denoted as $\mathrm{TT}_p(s,d,t_s)$, represents required duration to travel along the path p connecting s to d. Thus, the time-dependent fastest path $\mathrm{TDFP}(s,d,t_s)$ is lowest duration among all potential fastest path: $\mathrm{TDFP}(s,d,t\ ) = \min\{\mathrm{TT}_i(s,d,t\ )\}$.

4. PRE-COMPUTATION STAGE USING VORONOÏ

For the sake of speeding up the real-time path search, it is wise to execute a pre-computation to assess the overall network taking into consideration historical data. However, the efficiency depends mainly on the accuracy of the defined heuristic function. The main goal of a heuristic function is to reduce considerably the set of potential nodes to scan – belonging to the shortest path. The authors in (Demiryurek et al., 2011) suggest that the Lower Bound Graph itself is a reliable indicator of the potential nodes. Thus, the pre-computation stage consists of, on one hand, the breaking down of the network to non-overlapping partitions, on the other hand, the computation of the lower bound border-to-border, node-to-border, and border-to-node distance labels.

Regarding the Network partitioning, we use the Voronoï diagram that consists in dividing the space into a number of regions. This is done by specifying beforehand a set of seeds (or generator points) that match each region. A region is constructed by considering all points closer to one seed than to any other. The metrics can be defined according to ones needs. Okabe in (A. Okabe et al., 2012) summarizes the work done over the last decades on the network Voronoï diagrams that defines the metrics as weighted shortest path distances, to overcome the limitation of the planar Voronoï diagrams. Mabrouk in (A. Mabrouk et al., 2009) suggests a fuzzy network Voronoï with a fuzzy weighted metrics. We define the Fuzzy Voronoï Graph $\mathrm{F-VG}$ by breaking down the graph to sub-graphs such that each sub-graph contains the closest – in terms of fuzzy fastest path – vertices and edges to predefined Voronoï seeds (A. Mabrouk et al., 2009).

Executing Voronoï partitioning, we can deduce the lower bound travel-time, which is the lowest possible travel-time and never overestimate it, connecting border-to-border $\mathrm{LTT}(\mathrm{b}\ ,\mathrm{b_d})$, node-to-border $\mathrm{LTT}(\mathrm{s},\mathrm{b_s})$, and border-to-node $\mathrm{LTT}(\mathrm{b_d},\mathrm{d})$. Then, the heuristic function $\mathrm{h}(\mathrm{v})$ can be then expressed, according to (Demiryurek, et al., 2011), as

$$\mathrm{h}(\mathrm{s}) = \mathrm{LTT}(\mathrm{s},\mathrm{b}\ \ ) + \mathrm{LTT}(\mathrm{b}\ \ ,\mathrm{b_d}) + \mathrm{LTT}(\mathrm{b_d},\mathrm{d})$$

5. FUZZY COST

Fuzzy Logic represents a continuum degree of logic and it deals with approximate reasoning rather than crisp and fixed logic {true, false}. The asset of this approach is the similarity with real world logic. For instance, in a real-world scenario the travel-time cost of a road segment is not exact throughout a specific time unit, due to driving behaviors and traffic density. Subsequently, the use of approximate crisp values does not portray an appropriate degree of dynamics as much as fuzzy values do. The approach can be, then, extrapolated to all fuzzy attributes of the system, such as risk fluctuation.

The substitution of a Crisp Sets by Fuzzy Sets, which represents time-travel or risk, will eventually raise the issue of ranking Fuzzy Sets. Since this matter is less obvious and more complex comparing to crisp sets, that is Real Numbers. Several studies and propositions have been suggested since Fuzzy Sets introduction (Zadeh, 1965). We have proposed in (A. Boulmakoul et al., 2013) an original way for comparing fuzzy sets that takes advantage of the topological relationship of two triangles and the degree of inclusion in the MIN of two fuzzy sets.

6. MULTIAGENT SIMULATION SYSTEM

Multiagent System MAS is a system composed of multiple interacting computing elements, known as agents (Wooldridge, 2009). MAS are a recent subfield of computer science; they have only been studied since 1980 and gained widespread recognition since the mid-1990s (Wooldridge, 2009). The international interest in the field has expanded mostly to its intrinsic distributed computing property, which reduces the complexity and task processing time.

An agent is a computer system that is situated in some environment, and that is capable of autonomous action in this environment in order to meet its design objectives (Wooldridge et al., 1995). Agent takes sensory input from the environment, and produces as output actions that affect it. The interaction is usually an on-going, non-terminating one (Payne, 1971).

7. TRAFFIC SIMULATION

Two classes of simulation techniques are usually used in modeling traffic flows. Microscopic approaches consider discrete autonomous entities with complex set of rules governing their behavior. Macroscopic or Continuum methods allow the modeling of aggregated behavior of many vehicles. While the latter provide tools for large-scale traffic flows using fluid dynamics equations, the micro approach captures highly detailed information about each vehicle considering their intrinsic properties, in addition to the interaction with either other vehicles or the environment. Thus, it has been decided to use micro-simulation for DGT since we are interested by a detailed analysis. It generates discrete and pseudo-real data that allows accurate assessment and particular case studies. Various models have emerged approaching the microscopic traffic simulation, such DynaMIT, TRANSIMS, SUMO, and MATSim. In general, a traffic simulation system based on multi-agent architecture is designed in such a way that: Each vehicle is an agent; the network of roads is the environment; each agent is tied with a set of behaviors that manifest the responses of the agent to the environment changes; Agent acts autonomously.

8. HAZMAT SPACE TIME PATH DATA WAREHOUSE

The risk of transporting hazardous material is the result of a transport accident or incident on health and environment. Catastrophic accidents involving hazardous substances can occur suddenly anywhere at any moment. Given that consequences of an incident are often considerable, location based services purpose is to help against the immediate consequences of the disaster and the random nature of first aid. Several steps must be taken in near real time to avoid the damage. Aside, restrictive regulations based the training of staff (drivers) application of strict driving and traffic rules, an obligation tanks approval for vehicles signaling and hazardous products transported.
We propose in the following the use of a data warehouse in near time to help in case of an eventual risk, by sending alerts to people near to the disaster geographical location, to prohibit some roads and find backup routes. Several research efforts have been carried out on management of trajectories. Work given in (Spaccapietra et al., 2008) provides a semantic view of trajectory, which enables applications to associate whatever semantics they want with trajectories.

In the proposed schema of Figure 2, we used hierarchy documents that contain several related levels. In the following, we describe dimensions of the proposed space time path conceptual schema:
Vehicle dimension: contains information about the tracked vehicle like reference number, type of vehicle, traveled kilometers, capacity, mobile sensor type and also different signalization. The vehicle can carry several hazardous materials. This latter is characterized by ONU number and quantity, and it belongs to a determined class according to ADR. In addition, we also warehouse the type of HAZMAT packaging with its entire characteristic for a full Hazardous space-time paths analysis. The vehicle dimension is related to the driver entity with his history of training.

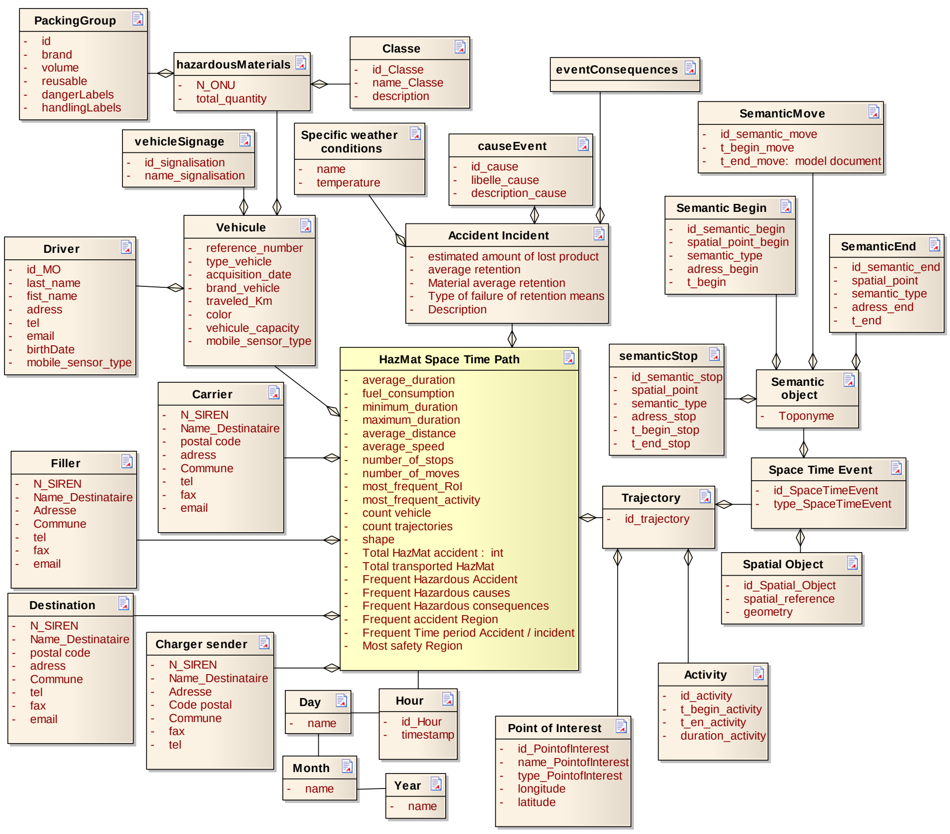

Figure 2. HAZMAT Space Time Path Data warehousing conceptual schema using composite documents.

Recipient dimension: presents the company that supports the Hazardous materials on arrival. Carrier dimension: presents the company that transports Hazardous materials with or without transport contract. Filler dimension: presents the company that fills dangerous materials in the tracking vehicle. Charger dimension: presents the company that load packaged Hazardous materials in a vehicle. Accident / Incident dimension: To minimize damage and facilitate analysis of accidents / incidents during the transport of materials, the following information is warehoused: the estimated amount of lost product, the average retention, the material retention means, type of failure of retention means and description of the event.

The near real time decision when an accident/ incident occur in a hazardous space-time path requires a full knowledge about Specific weather conditions, Cause of the event, and Consequences of the event. Event ontology has already been proven useful in a wide range of contexts, due to its simplicity and usability. The SHOE General Ontology defines an event as something that happens at a given place and time. Space Time Event: presents an event as an occurrence that happens in a small space and lasts a short time. From spatial point of view, it is a composition of Spatial Object. Each spatial object is characterized by a spatial reference and geometry to model a raw trajectory. From semantic point of view, a semantic object is characterized by a toponyme, and linked to the semantic begin move end and stop which respectively contains semantic information, time of begin and of the end for a given begin, move stop and end. To analyze the space time path data warehouse, the schema considers the presentation of spatio-temporal data and activities for each event. As concerning the temporal aspects, it is detailed with others dimensions tables like Hour, Day, Month, Year and Time Period.

Additionally, HAZMAT space-time path data warehouse schema presents HAZMAT trajectories in both spatial and temporal contexts based on Region of Interest and activities. The spatial neighborhood is presented using Point of Interest dimension characterized by longitude, latitude and name. Area of interests has a shape and a surface. City region of interests to present a city as a region of interests, it contains about the city region of interest name, surface, first language and religion. In other applications domain, region of interests could be presented as a Voronoi network of interest or a modal network. The system architecture of data warehouses is traditionally associated with the relational model as operational database and SQL as a query language. The challenge of analyzing of the massive HAZMAT space-time path data warehouse in near real time requires the use of a fast and scalable solution to bear the burden of voluminous data.

In this work, we provide a NoSQL database for the storage of HAZMAT space-time path data warehouse. All given components are integrated into an interoperable software infrastructure respecting intelligent transport systems architecture. This infrastructure is distributed and based on a service-oriented architecture. It is also scalable by integration of MongoDB with Hadoop for large-scale distributed data processing. In this work, we also give an assessment of the performance, scalability and fault-tolerance of using MongoDB with Hadoop, towards the goal of identifying the right architecture and software environment for HAZMAT spatio-temporal data analytics. In the following, we give a brief description about MongoDB and Hadoop:

MONGODB is an open-source document database, and the leading NoSQL database. MongoDB (from "humongous") is an open-source document database and the leading NoSQL database developed by 10gen in 2009 (MongoDB, 2013). In general, document-oriented (e.g. MongoDB) are most directly relevant to business intelligent because of their more flexible and extensive search and retrieval functionality. To scale its performance on a cluster of servers, MongoDB uses a technique called sharding, which is the process of splitting the data evenly across the cluster to parallelize access.

HADOOP (Alex, 2012) is the Apache Software Foundation top-level project that provides both distributed storage and computational capabilities. The Hadoop project provides and supports the development of open source software that supplies a framework for the development of highly scalable distributed computing applications. The Hadoop Core project provides the basic services for building a cloud computing environment with commodity hardware, and the APIs for developing software that will run on that cloud. The two fundamental pieces of Hadoop Core are the MapReduce framework, the cloud-computing environment, and the Hadoop Distributed File System (HDFS).

9. CONCLUSION

In this paper we have proposed calculating safer time-dependent routes, spatial analysis of Voronoi network and establishment of a decisional database to capture HAZMAT shipments and occurring incident or accidents. The proposed system can be exploited in different applications domains and is able to handle in near real time GeoStream amount of spatio-temporal data of hazardous trajectories system from different moving objects and analyzing them in a scalable, fast and agile way.

Decision Makers can mine HAZMAT space-time path data warehouse in MongoDB data-base using Hadoop framework and its MapReduce paradigm to benefit from the maximum of performance and scalability.

10. REFERENCES

Basavaraju A., Doddigarla S., Naidu N., Malgatti S. (2014). Vehicle Density Sensor System To Manage Traffic. IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

Holmes A. (2012). Hadoop in Practice. Manning Publications Co.

Aydin, G., Qi, Z., Pierce, M., Fox, G., Bock, Y. (2007). Architecture, performance, and scalability of a real-time global positioning system data grid. Physics of the Earth and Planetary Interiors 163(1-4) pages 347-359.

Benza, M., Bersani, C., D'Inc, M., Roncoli, C., Trotta, A., Pizzorni, D., Briata, S., Ridol, R. (2012). Intelligent transport systems (its) applications on dangerous good transport on road in italy. In: IEEE SOSE 2012 7th International Conference On System Of Systems Engineering. Volume Risks and Transportation, pages 223-228. IEEE.

Boulmakoul, A., Laurini, R., Zeitouni, F. (2001). Spatial monitoring and routing system for the transportation of hazardous materials. In: Environmental information systems in industry and public administration, IGI Publishing 227-236.

Boulmakoul, A., Laarabi, M.H., Sacile, R., Garbolino, E. (2013). Ranking triangular fuzzy numbers using fuzzy set inclusion index, in Fuzzy Logic and Applications, ser. Lecture Notes in Computer Science, F. Masulli, G. Pasi, and R. Yager, Eds. Springer International Publishing, vol. 8256, pages 100–108. Available: http://dx.doi.org/10.1007/978-3-319-03200-9 11

Ehlert, P. A., Rothkrantz, L.J. (2001). Microscopic traffic simulation with reactive driving agents. In Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE, pages 860–865.IEEE.

Garbolino, E., Sacile, R., Olampi, S., Bersani, C., Tomasoni, A., Alexandre, N., Trasforini, E., Benza, M., Giglio, D. (2007). Spatial decision support system prototype for assessing road hazmat accident impacts on the population in a dense urban area: a case study of the city of nice, french riviera. In: Proceedings of the 8th International Conference of Chemical and Process Engineering 2007, in Ischia, pages 413-418.

Laarabi, M.H., Boulmakoul, A., Mabrouk, A., Sacile, R., Garbolino, E. (2014, May). Real-Time Fastest Path Algorithm using Bidirectional Point-to-Point Search on a Fuzzy Time-Dependent Transportation Network. In 3rd International Conference on Advanced Logistics and Transport, 2014. IEEE.

Laarabi, M.H., Roncoli, C., Sacile, R., Boulmakoul, A., Garbolino, E. (2013, April). An overview of a multiagent-based simulation system for dynamic management of risk related to dangerous goods transport. In Systems Conference (SysCon), 2013 IEEE International, pages 830-835. IEEE.

Mabrouk, A., Boulmakoul, A., Bielli, M. (2009). Fuzzy spatial network voronoi diagram: a spatial decision support for transportation planning, International Journal of Services Sciences, vol. 2, no. 3, pages 265–280.

Ji, B., Ji, S., Zhang, C. (2011). Gprs-based data real-time transmission system of water-quality monitoring. In: Electronics, Communications and Control (ICECC), 2011 International Conference on, IEEE, pages 2072-2075.

Okabe, A., Sugihara, K. (2012). Network Voronoi Diagrams. John Wiley & Sons, Ltd. pages 81–100. Available: http://dx.doi.org/10.1002/9781119967101.ch4

MongoDB 10gen. (2013). Available from: http://www.mongodb.org.

Payne, H. J. (1971). Models of freeway traffic and control. Mathematical models of public systems.

Spaccapietra, S., Parent, C., Damiani, M.D., Macedo, J.A., Porto, F., Vangenot, C. (2008). A Conceptual view on trajectories. Data and Knowledge Engineering, pages 26–146.

Tomasoni, A.M., Garbolino, E., Rovatti, M., Sacile, R. (2010). Risk evaluation of real-time accident scenarios in the transport of hazardous material on road. Management of Environmental Quality: An International Journal 21(5) 69-711.

Demiryurek U., Banaei-Kashani F., Shahabi C., Ranganathan A. (2011). Online computation of fastest path in time-dependent spatial networks, in Advances in Spatial and Temporal Databases. Springer, pages 92–111.

Wooldridge, M. (2009). An introduction to multiagent systems. Wiley.com.

Wooldridge, M., Jennings, N.R. (1995). Intelligent agents: Theory and practice. Knowledge engineering review 10(2), pages 115–152.

Xu, Z., Fang, Z. (2008). A clustered real-time remote monitoring system for out-of-hospital cardiac patients. In: BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on. Volume 1., IEEE 610-614.

Zadeh, L. A. (1965). Fuzzy sets. Information and control 8(3), pages 338–353.